Requested Patent: JP2000315200A

Title: ;

Abstracted Patent: JP2000315200 ;

Publication Date: 2000-11-14 ;

Inventor(s): ;

Applicant(s): ;

Application Number: JP19990309760 19990924 ;

Priority Number(s):

US19980101656P 19980924; US19990398248 19990917 ;

IPC Classification: G06F15/177 ; G06F13/00 ;

Equivalents:

ABSTRACT:

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2000－315200
（P2000－315200A）
(43)公開日　平成12年11月14日(2000.11.14)

(51)Int.Cl.'		識別記号	FI		テーマコード'(参考)
G06F	15/177	674	G06F	15/177	674A
	13/00	357		13/00	357Z

審査請求　未請求　請求項の数16　OL　外国語出願　（全 39 頁）

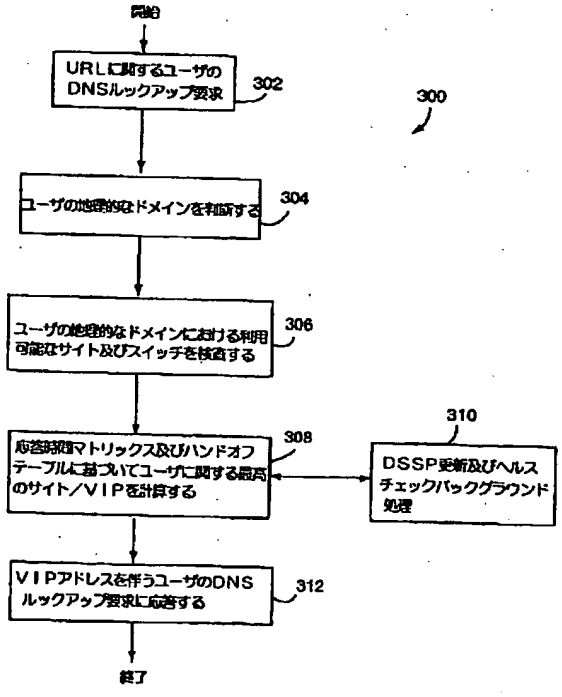
(21)出願番号　特願平11－309760
(22)出願日　平成11年９月24日(1999.9.24)
(31)優先権主張番号　60／101656
(32)優先日　平成10年９月24日(1998.9.24)
(33)優先権主張国　米国（US）
(31)優先権主張番号　09／398248
(32)優先日　平成11年９月17日(1999.9.17)
(33)優先権主張国　米国（US）

(71)出願人　599152522
アルテオン　ウェブ　システムズ　インコーポレイテッド
アメリカ合衆国　カリフォルニア州　95119　サン　ホセ　サン　イグナシオ　アベニュー　6351
(72)発明者　ディヴィッド　ピー　ローガン
アメリカ合衆国　カリフォルニア州　95138　サン　ホセ　ジェントリー　オークス　プレイス　6749
(74)代理人　100059959
弁理士　中村　稔　（外９名）

最終頁に続く

(54)【発明の名称】　分散されたロード平衡インターネットサーバ

(57)【要約】　（修正有）
【課題】　分散されたロード平衡インターネットサーバを提供する。
【解決手段】　ピアハンドオフプロセスにおいて、スイッチは特定のドメインネームのためのドメインネームサーバルックアップリクエストを受け取る。スイッチはソースIPアドレスを調べ、ユーザのIPアドレスを調べ、ユーザに地図上で近いサーバサイトがないか調べる。スイッチはドメインに対応する順序付けられたハンドオフテーブルを調べる。スイッチは、（ａ）ドメインネームサーバリクエストソースと対照的な遠隔サーバ位置、（ｂ）遠隔サーバの負荷、及び（ｃ）前のハンドオフを経験した遠隔サーバをベースとする次の遠隔サーバ（または独自の仮のインターネットプロトコルアドレス）を選択する。スイッチは次いで、順序付けられたリストにおけるIPアドレスとともにクライアントドメインネームにドメインネームサーバレスポンスを送り返す。



【特許請求の範囲】

【請求項１】ドメインホストネームに関する特定のユニフォームリソースロケータ（ＵＲＬ）の変換に関してネットワークユーザから、数値のインターネットプロトコル（ＩＰ）アドレスに対するＤＮＳルックアップ要求を受信するドメインネームシステム（ＤＮＳ）サーバを有し、前記ネットワークユーザは、ユーザＩＰアドレスから識別されうる特定の地理的な領域に存在し、

地理的に様々であり、前記ネットワークユーザにアクセス可能である複数のウェブサーバサイトを有し、各々が、前記特定のＵＲＬに関するそのウェブベースのコンテンツ及びサービスを別に複製し、

複数のウェブサーバサイトの各々のヘルス及び応答パフォーマンスを監視し、個々のアクセス可能性及び地理的ロケーションに関する複数のウェブサーバサイトのうちのリストを保守するポリシーマネージャと、

前記ネットワークユーザからの前記ＤＮＳルックアップ要求を受信するために接続され、かかるＤＮＳルックアップ要求に応答するように複数のウェブサーバサイトのうちの好ましい１つに関するポリシーマネージャを調べるために接続され、更に、複数のウェブサーバサイトのうちの前記好ましい１つのＩＰアドレスを備える前記ネットワークユーザを提供するために接続されるＩＰアドレスコンバータに対するＤＮＳクエリーと、を有する、多くのクライアントに冗長に配送されるウェブベースのコンテンツ及びサービスを提供する、分散されたロード平衡インターネットサーバシステム。

【請求項２】複数のウェブサーバサイトの各々が仮想ＩＰアドレス（ＶＩＰ）に対応し、世界中の異なる場所で物理的に位置決めされる、請求項１に記載のシステム。

【請求項３】複数のウェブサーバサイトの各々が、その他のものをオフ・ロードすることができ、異なった地理的ロケーションを有する多くの同時のネットワークユーザの要求を満たすように並列に作動することができる、請求項１に記載のシステム。

【請求項４】ＤＮＳクエリー－ＩＰアドレスコンバータは、システム・ワイド・ロードが複数のウェブサーバサイトの各々の間で平衡がとられるように作動する、請求項１に記載のシステム。

【請求項５】ポリシーマネージャが、応答時間マトリックスと、前記リストを保守するハンドオフ・テーブルとを更に含む、請求項１に記載のシステム。

【請求項６】ポリシーマネージャが、最悪のパフォーマンスのウェブサーバサイトに対する最高のパフォーマンスに対応するハンド・オフ重みインデックスと、サーバスイッチのオーダされたハンド・オフ・リストにおける相対的な位置によって増やされた静的に構成された重みとを有する、請求項５に記載のシステム。

【請求項７】ポリシーマネージャが更に、インターネットトポロジーアウェアネスと、ヘルス、ロード、周期的又は所定のイベントによってトリガーがかけられたときのいずれかにおけるウェブサーバサイトの間のスループット情報とを交換することができる分散されたＳＬＢステートプロトコルと、を有する、請求項５に記載のシステム。

【請求項８】複数のウェブサーバサイトが、互いのウェブサーバサイトによる複製に関する全てのウェブコンテンツ及びサービスを提供するメインコンテンツサイトを含む、請求項１に記載のシステム。

【請求項９】ドメインホストネームに関する特定のユニフォームリソースロケータ（ＵＲＬ）の変換に関してネットワークユーザから、数値のインターネットプロトコル（ＩＰ）アドレスに対するＤＮＳルックアップ要求をドメインネームシステム（ＤＮＳ）サーバで受信し、前記ネットワークユーザは、ユーザＩＰアドレスから識別されうる特定の地理的な領域に存在し、

前記ネットワークユーザにアクセス可能である地理的に様々なロケーションで、複数のウェブサーバサイトを配置し、各ウェブサーバサイトが、前記特定のＵＲＬに関するそのウェブベースのコンテンツ及びサービスを別に複製し、

複数のウェブサーバサイトの各々のヘルス及び応答パフォーマンスをポリシーマネージャで監視し、個々のアクセス可能性及び地理的ロケーションに関する複数のウェブサーバサイトのうちのリストを保守し、

前記ネットワークユーザからの前記ＤＮＳルックアップ要求の受信に応答してＤＮＳクエリーをＩＰアドレスに変換し、かかるＤＮＳルックアップ要求に応答するように複数のウェブサーバサイトのうちの好ましい１つに関する前記ポリシーマネージャを調べるために接続し、更に、複数のウェブサーバサイトのうちの前記好ましい１つのＩＰアドレスを備える前記ネットワークユーザを提供するために接続する、ステップを有する、

単一のＤＮＳルックアップ要求に応答してロード平衡冗長サイトから多くのクライアントにウェブベースのコンテンツ及びサービスを提供する方法。

【請求項１０】複数のウェブサーバサイトの各々が仮想ＩＰアドレス（ＶＩＰ）に対応し、世界中の異なる場所で物理的に位置決めされるように、複数のウェブサーバサイトを位置決めする、請求項９に記載の方法。

【請求項１１】複数のウェブサーバサイトの各々が、その他のものをオフ・ロードすることができ、異なった地理的ロケーションを有する多くの同時のネットワークユーザの要求を満たすように並列に作動することができるように、複数のウェブサーバサイトを配置する、請求項９に記載の方法。

【請求項１２】システム・ワイド・ロードが複数のウェブサーバサイトの各々の間で平衡がとられるようにＤＮＳクエリー－ＩＰアドレスコンバータが作動するように、変換する請求項９に記載の方法。

【請求項１３】ポリシーマネージャが、応答時間マトリックスと、前記リストを保守するハンドオフ・テーブルとを更に含むように監視する、請求項９に記載の方法。
【請求項１４】前記ポリシーマネージャが、最悪のパフォーマンスのウェブサーバサイトに対する最高のパフォーマンスに対応するハンド・オフ重みインデックスと、サーバスイッチのオーダされたハンド・オフ・リストにおける相対的な位置によって増やされた静的に構成された重みとを更に有するように監視する、請求項１３に記載の方法。
【請求項１５】前記ポリシーマネージャが更に、インターネットトポロジーアウェアネスと、ヘルス、ロード、周期的又は所定のイベントによってトリガーがかけられたときのいずれかにおけるウェブサーバサイトの間のスループット情報とを交換することができる分散されたＳＬＢステートプロトコルと、を有するように監視する、請求項１３に記載の方法。
【請求項１６】前記複数のウェブサーバサイトが、互いのウェブサーバサイトによる複製に関する全てのウェブコンテンツ及びサービスを提供するメインコンテンツサイトを含むように配置する、請求項９に記載の方法。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は一般に、コンピュータデータネットワーク装置及び方法に関し、より詳細には、ドメインネームをドメインネームサーバ装置におけるIPアドレスに変換することを制御することによりローディングを分配されたネットワークサーバに対して平均化することに関する。
【０００２】
【従来の技術】ワールドワイドウェブ（WWW）、特にインターネットは急速に、ビジネスにおいて製品を売り、顧客や供給者とコミュニケーションをとる本質的な方法ととなっている。今、インターネットを「ミッション・クリティカル(mission critical)ビジネスデリバリーインフラストラクチャー」と呼ぶ者もいる。結果として、インターネットサーバ及びいわゆる「イントラネット」サーバは以前にも増してすさまじく働いている。今支持せねばならないクライアントサーバの数多くは劇的に増加した。イントラネットサーバは、外部カウンターパートイントラネットサーバが何万もの同時クライアントコネクションを支持可能である時、何百もの同時クライアントリクエストを提供しなければならない。
【０００３】クライアントは迅速な応答及び1週間に７日、１日２４時間（「７×２４」）の稼働率を要求し、かつ期待する。ミッション・クリティカルウェブ・コンピューティングインフラストラクチャーは力学的にサーバ容量を基準化して、集合したクライアントデマンドに合わせ、連続的なサービス可用性を更に保証するようにしなければならない。唯一の方法は、いくつかのサーバ上で各アプリケーションを実行し、次いで連続的に種々のサーバ、例えば「サーバロードバランシング」上でクライアントローディングを平均化することである。
【０００４】サーバロードバランサーは、アプリケーション層セッションを識別し、管理するように層３及び層４パケットヘッダーにおける情報を使用する。例えば、TCP又はUDPポートナンバー、TCPアプリケーションセッション及びIPソースの始まりと終わりを示すSYN/FINビット、及びデスティネーションアドレスである。
【０００５】伝統的なサーバロードバランサーはリミテッドパフォーマンス及び連結性を備えたPCベースのソフトウェア製品である。トラフィック容量及びサーバ母集団における急速な成長により、性能、連結性、弾性、及び節約における大規模な改良の命令を与えるスイッチ統合サーバロードバランサーの新世代を引き起こす。
【０００６】スイッチベースのサーバロードバランサーの新世代は、キャッシュに対するリダイレクション、複数のファイアウォールに対するロードバランシングトラフィック、パケットフィルタリング、及び帯域幅管理のような複数のウェブインフラストラクチャー機能及び多層スイッチングを備えたロードバランシングアプリケーションサーバを合同する。
【０００７】アルテオン（Alteon）ウェブシステムは「サーバスイッチ」という用語を造りだして、サーバファームをフロントエンドする装置の新しい分類を表し、全てのミッションクリティカルインターネット／イントラネットインフラストラクチャーにおけるサーバ関連トラフィック管理を提供した。サーバスイッチは、グループがネットワークに対する１つのサーバであるかのようにする一方で、通常のアプリケーション（あるいは１組のアプリケーション）を実行するサーバのグループにわたってアプリケーションロードを力学的に分配する。同じ内容にアクセスするウェブサーバの数多くは、通常のアプリケーション又は１組のアプリケーションを支持するサーバのグループであるHTTPハントグループに必然的に結合される。ハントグループはクライアントに対する「仮の」HTTPサービスを提供する。クライアントは多くの本物のサーバがこのサービスの提供に参加していることに気付いていない。クライアントは、本物のサーバをフロントエンドするサーバスイッチに属する仮のサービスアドレスを使用するサービスにアクセスする。連結において仮のサービスのために到着することを要求する時、サーバスイッチはこれらの要求をサーバの可用性、ロードハンドリング能力、および現行のロードの知識によるハントグループにおける本物のサーバの１つに渡す。
【０００８】このようにして、複数のサーバはシステムの使用者によって要求されたアプリケーション処理容量の総量を達成するのに使用される。各新サーバは容量をアプリケーショに有用なたまった処理電力に加える。

【０００９】同様に重要であるが、サーバが故障又はメンテナンスオペレーションのいずれかによりサービスが止まった時、残った健全なサーバは使用者に近くされた衝撃を多少あるいは全く与えずにロードをピックアップする。これを達成するためには、サーバスイッチはクライアントロードを分配する全てのサーバ及び各アプリケーションの健康状態を連続的に監視しなければならない。サーバスイッチはまた、完全なシステム余剰のためのホットスタンバイ形態を支持しなければならない。

【００１０】サーバロードバランシングのキーの部分は、セッション管理である。一旦セッションリクエストが本物のサーバに割り当てられると、サーバスイッチは、セッションと関連した、全ての継続的なパケットを認識しなければならない。これらのパケットはクライアントが各セッションの存続期間、同一の物理的サーバと関連して継続するのを確かめるように、適切に処理され、かつフォワードされる。

【００１１】サーバスイッチはまた、物理的サーバへの連結の結合が取り除かれる時のセッションの完了を監視する。これは次回、最も有用なサーバに連結するのが好ましいクライアントが連結し、各クライアントに最も可能な限りのサービスを提供することを保証する。アプリケーションによって継続的な連結がFTP制御及びデータ連結、SSL（Secure Sockets Layer）、及びマルチページフォーム及びサーチエンジンに使用される持続的なHTTPのような同一の物理的サーバにフォワードされることを必要とするならば、特別なメカニズムがアドミニストレータにより呼び出される。

【００１２】サーバロードバランシングから利益を得る環境はウェブホスティングサービス、オンラインサービスプロバイダー、及び高有用性要求物を備えたコーポレートデータセンターを含む。理論上は、サーバロードバランシングは通常の内容がサーバのグループにわたって有用である、TCPベースの、又はUDPベースのアプリケーションを支持するように使用される。実際問題として、ウェブサーバ、FTPサーバ、ドメインネームサーバサーバ、及びRADIUSサーバのようなインターネット／イントラネットアプリケーションを支持するサーバは、好ましくはサーバロードバランシングによって高成長及びウェブ指向のトラフィックの予測できない容量を支持することを最初に利用する。

【００１３】ウェブページの大多数は読取専用情報を含む。これはウェブホスティング環境がサーバロードバランシングにとって理想的であるようにする。ウェブホスト及びオンラインサービスプロバイダーは今日典型的に複数のHTTP、FTP及び他のアプリケーションサーバを展開し、ロードを静的に分配し、あるいはもっと一般的にはラウンドロビンドメインネームサーバを経て分配する。両方の方法は、故障許容がなく、高度の管理を必要とするため、望ましくない。サーバロードバランシングはルトイン高有用性支持との複数のサーバの即応型使用を可能にする。

【００１４】今日の多くのクラスタ化システムは超越したフェイルオーバー能力を提供するが、ロードバランシング支持を提供しない。いくつかのシステムはまた、クラスタに関係する数多くのサーバを制限する。これらの制約はクラスタ化の解のスケーラビリティに衝撃を与える。サーバロードバランシングはロード共有ハントグループにサーバを柔軟に結合することを可能にする。また、冗長サーバがロードを共有するようにすることによってサーバユーティライゼイション効率を改善する。

【００１５】たいてい、今日のサーバ環境はマルチベンダ及びマルチOSである。今日の一般的なクラスタ化の解は単一のベンダからのサーバ又は単一のオペレーティングシステムを実行するサーバに限定される。サーバスイッチ上のサーバロードバランシングはTCP及びUDPアプリケーションを支持する異質のサーバがロード共有クラスタにゆるく結合して、サーバインベストメントリターンを最大化する事を可能にする。

【００１６】

【課題を解決するための手段】URLドメインネームに応答してクライアントにウェブページを提供する実際のインターネットウェブサイトは、同一のデータ記憶装置を夫々が有する、分配された多くのサイトのリストから自動的にかつ即応して選択される。ピアハンドオフプロセスにおいて、スイッチは特定のドメインネームのためのドメインネームサーバルックアップリクエストを受け取る。スイッチはドメインネームサーバリクエストのためのソースIPアドレスを調べ、ユーザのIPアドレスを調べ、ユーザに地図上で近いサーバサイトがないか決定する。スイッチはドメインに対応する順序付けられたハンドオフテーブルを調べる。スイッチは、（ａ）ドメインネームサーバリクエストソースと対照的な遠隔サーバ位置、（ｂ）遠隔サーバの重量、及び（ｃ）前のハンドオフを経験した遠隔サーバをベースとする次の遠隔サーバ（または独自の仮のインターネットプロトコルアドレス）を選択する。スイッチは次いで、順序付けられたリストにおけるIPアドレスとともにクライアントドメインネームにドメインネームサーバレスポンスを送り返す。

【００１７】

【発明の実施の形態】図１は、本発明の分散サーバ負荷均衡(distributed-server load-balancing)システムの実施形態を表わし、ここに一般的な参照符号１００により示される。分散サーバ負荷均衡システム１００により、ウェブベースのコンテンツ及びサービスは、クライアント”Ｚ”１０２により表わされる多くのクライアントに、多くの独立したウェブーサーバサイトからインターネット１０４を通して冗長に送られる。クライアント１０２がウェブブラウザプログラムをロードしたとき、例えばwww.alteon.com/products/index.htmlのようなユ

ニフォームリソースロケーション（ＵＲＬ）に入る。

【００１８】インターネット１０４上で使用されるＩＰアドレスは３２ビット長であるが、ほとんどのユーザは、彼らが加入するホストの数字のアドレスを覚えていない。その代わりに人々は、ホストネームによる方がより快適である。ほとんどのＩＰホストは、そのため、数字のＩＰアドレス及びネームの両方を有している。これは人々にとっては便利である一方、ルーティングの目的のためには、そのネームは数字のアドレスに逆に転換されなくてはならない。インターネットホストは、トップレベルドメイン（top-level domain,ＴＬＤ）、ドメイン及び（必要に応じ）サブドメイン、及びホストネームを含む、階層状ネーミング構造(hierarchical naming structure)を使用している。ＩＰ－アドレス空間、及び全てのＴＣＰ／ＩＰ関連の番号は、インターネットアサインドナンバーズオーソリティ（Internet Assigned Numbers Authority,ＩＡＮＡ）により割り当てられ、維持されている。ドメインネームは、ＴＬＤネーミングオーソリティにより割り当てられる；１９９８年４月まで、インターネットネットワークインフォメーションセンター（Internet Network Information Center, ＩｎｔｅｒＮＩＣ）は、これらのネームの全般の機関であったし、世界中のＮＩＣが米国以外のドメインを扱っていた。ＩｎｔｅｒＮＩＣは、インターネット上のホストネームとＩＰ－アドレスとを両立させる分散データベースである、ドメインネームシステム（ＤＮＳ）の全般の調整及び管理にも責任があった。

【００１９】クライアント－Ｚ　１０２では、ドメインネームサーバの”getByHostname”照会が、ローカルドメインネームサーバに実際に発行され、www.alteon.comで使用するために登録された数字のインターネットプロトコルアドレス（ＩＰ－アドレス）を要求する。それぞれのローカルドメインネームサーバは、特定のドメインネーム及びホストにサービスを提供するホストのためのＩＰ－アドレスを、それが既に知っているかどうかを確認する。それは、この情報を先立って必要とすること及びローカルプライベートキャッシュメモリ中にそれが発見した答を記憶させることにより、このことを知ることができよう。もしローカルドメインネームサーバが、要求されたＵＲＬドメインネームに対するホストネームＩＰ－アドレスを知らないなら、それは、ＤＮＳ階層中でより高位のドメインネームサーバへ反復する照会を実行するであろう。そのようなドメインネームサーバの照会は、より高いレベルのドメインネームサーバによって答えられるか、あるいはその要求は、分散サーバネットワークスイッチサイト１０６、１０８、又は１１０の内の１つに、最後には表面に浮かんでくるかのどちらかであろう。

【００２０】ＩＰ－アドレスは、ルーティングの目的のために階層状であり、さらに２つのサブフィールドに分けられている。ネットワーク識別子(Network Identifier,ＮＥＴ＿ＩＤ)サブフィールドは、インターネットに接続されたＴＣＰ／ＩＰサブネットワークを識別する。ＮＥＴ＿ＩＤは、ネットワーク間の高レベルのルーティングのために使用され、カントリーコード、シティコード又はエリアコードとほぼ同じ方法が電話ネットワーク中で使用される。ホスト識別子（ＨＯＳＴ＿ＩＤ）サブフィールドは、サブネットワーク内の特定のホストを指示する。

【００２１】ほとんどのＩＰホストは、数字のＩＰ－アドレス及びネームを通常有する。ネームは、人々のための便宜として提供されるが、そのようなネームは、ルーティングの目的のためには数字のアドレスに逆に転換されなくてはならない。インターネットホストは、トップレベルドメイン（top-level domain,ＴＬＤ）、ドメイン及び（必要に応じ）サブドメイン、及びホストネームを含む、階層状ネーミング構造(hierarchical naming structure)を使用している。分散サーバネットワークスイッチ１０６、１０８及び１１０は、分散サイトとして編成され、ここでそれぞれは、例えばwww.alteon.comのようなサブドメインに対する正当ネームサーバ(Authoritative Name Server)として働く。それぞれのそのような分散サイトは、www.alteon.comに対応するＩＰ－アドレス識別を有するドメインネームサーバの照会に応答することができる。

【００２２】ＴＣＰ／ＩＰプロトコルの一続きは、ＯＳＩトランスポート(OSI Transport)及びセッションレイヤ(Session Layer)におおよそ対応する２つのプロトコルを含む。これらのプロトコルは、トランスミッションコントロールプロトコル(Transmission Control Protocol)及びユーザデータグラムプロトコル(User Datagram Protocol,ＵＤＰ)と呼ばれる。個々のアプリケーションは、ＴＣＰ／ＵＤＰメッセージ中のポート識別子により表わされる。ポート識別子及びＩＰ－接続は、一緒にソケットを形成する。接続のサーバ側上の周知のポート番号は、ポート－２０（ＦＴＰデータ転送）、ポート－２１（ＦＴＰコントロール）、ポート－２３（Telnet,電話ネットワーク）、ポート－２５（ＳＭＴＰ）、ポート－４３（whois,誰か）、ポート－７０（Gopher,ゴーファー）、ポート－７９（finger,指）、及びポート－８０（ＨＴＴＰ）を含む。

【００２３】図解の目的のため、分散サーバスイッチ１０８は、クライアント１０２により発せられたドメインネームサーバ照会を受け取ると仮定する。本発明の実施形態では、分散サーバスイッチ１０８は、仮想－ＩＰ（ＶＩＰ）を表わすＩＰ－アドレスの一組を返すであろう。例えば、分散サーバスイッチ１０８は、ＵＲＬ照会に、どれもが単一のＵＲＬに関連するウェブベースのコンテンツ及びサービスの要求を満足する”192.168.13.20”、”162.113.25.28”及び”172.176.110.10”を含む

ＩＰ－アドレスの一組でもって応答することができよう。これらの数個のＩＰ－アドレスのそれぞれは、例えば、分散サーバスイッチ１０６及び１１０によって表わされるような、地理的に異なったサーバに存在する。クライアント１０２は、そのような応答を、それのローカルドメインネームサーバ経由で受け取るであろう。そしてクライアント１０２は、これらのＩＰ－アドレスを使用し、例えば分散サーバスイッチ１０６で実際に働いているＶＩＰ－アドレスである”192.168.13.20”へのＴＣＰポート８０接続を開くことができる。クライアント１０２は、これが単なるＶＩＰであることを知らず、またスイッチ１０６に存在する”192.168.13.10”の実際のＩＰアドレスを無視することができる。その後、www.alteon.comウェブサイトを有するクライアント１０２により発せられたトラフィックは、分散サーバスイッチ１０６により扱われ、他の可能性のあるスイッチ１０８及び１１０からオフロード(off-load)される。

【００２４】それぞれのスイッチ１０６、１０８及び１１０のためのＶＩＰの設定(set up)は、任意の１つへの要求(request)がクライアント１０２に与えられている同じデータに帰するように、それぞれ同じコンテンツ及びアプリケーションへのクライアントのアクセスを可能にしなければならない。そのため、利用可能なリソースをユーザが必要とするサービスに分配する、ある手段を確立する必要がある。そのような手段において考慮すべき要因は、関連する個々の分散サーバのＶＩＰの健全性、基本的なインターネットアサインドナンバーオーソリティ（ＩＡＮＡ）が登録したクライアント及びサーバの位置、及び現時に測定された応答時間及びスループットに従って利用可能なサーバのリストを含む。最も健全で、より近くに位置し、良好な応答時間及びスループットを示すようなサーバは、それらにより多くのトラフィックが向けられるべきである。それらの対応するＶＩＰに応答することにより、非常にしばしば、このことは実行される。

【００２５】ＤＮＳは、インターネット上の各ドメインに対するホストネーム及びＩＰ－アドレス情報の普通の分散データベースである。各ドメインに対し、単一の正当ネームサーバ(authoritative name server)がある。やく１２個のルートサーバ(root server)が、こららの正当ネームサーバのすべてのリストを有する。ホストによりＤＮＳに要求(request)がなされたとき、その要求はローカルネームサーバに行く。もし、ローカルネームサーバのところでの情報が不十分あれば、正当ネームサーバを見つけるための要求がルートへなされ、また情報の要求(information request)は、そのネームサーバへ送られる。ネームサーバは、以下のタイプの情報を含む。

【００２６】Ａ－レコード：アドレスレコードは、ホストネームをＩＰ－アドレスにマッピングする。

【００２７】ＰＴＲ－レコード：ポインタレコードは、ＩＰ－アドレスをホストネームにマッピングする。

【００２８】ＮＳ－レコード：ネームサーバレコードは、所定のドメインに対する正当ネームサーバを一覧にして示す。

【００２９】ＭＸ－レコード：メール交換レコードは、所定のドメインに対するメールサーバを一覧にして示す。

【００３０】クライアント１０２が失敗を突然経験するように指摘されたか、又はオーバーロードであるサーバスイッチが１０６，１０８又は１１０ならば、それは「”HTTP redirect”」を発行する。従って、クライアント１０２は、異なるサーバスイッチ１０６、１０８又は１１０に向かうように命令される。「HTTP Request」が、最大通信（”MaxConns”）又はもはやいなかるヘルスリアルサーバもないＶＩＰで到着するとき、「”HTTP redirect”」は発行される。

【００３１】図１の分散サーバロード平衡システム１００は、ＶＩＰサイトに関するＤＮＳ要求に応答するようにドメインネームサーバを使用する。”www.alteon.com”の例は、Alteonウェブ分散サーバに関する同じコンテンツに対するアクセスを備える米国を介して散乱した種々のＶＩＰを示す。スイッチが、ＶＩＰと関係して”www.alteon.com”を解決するためにドメインネームサーバName Requestを受信するとき、次のコンテンツ要求に応答するために「ベストサイト」を適合する適当なドメインネームサーバレスポンスで応答する。例えば、かかるベストサイトは、ユーザの最大の数で、最小の遅延を負荷するものを表す。最小のコストであるものとして応答するベストサイトを構成するような、他の標準も可能である。

【００３２】サイトの稼働状態及びスループットの測定は、全ての他のピアリモートサイト１０６、１０８、及び１１０で「Ｌ４ヘルスーチェッキング」（オプションとして内容確認で）の間に得られる。このようなものは、アプリケーションの使用可能性の状態を決定し、また、各サイトのスループット性能を決定するのに使用される。

【００３３】稼働状態、負荷、スループット情報を、サイト間で、周期的にもしくは所定の事象によって引き起こされる時に交換することができる分散ＳＬＢ状態プロトコルが、使用される。インターネットトポロジーアウェアネスが本発明の実施例に含まれるのが好ましい。

【００３４】インターネットトポロジーアウェアネスに関し、ＤＮＳ／ＨＴＴＰハンドーオフのために使用される特定のスイッチが、要求のソース＿ＩＰを調べ、世界中のＩＰ－アドレス空間を割り当てられたＩＡＮＡに基づいて「最良の」サーバーで応答する。外部の「加入者データーベース」に、登録されたユーザーネットワークがどこに位置するかを記述する必要な両の細目を提供す

るように要求しても良い。この情報を、インターネット割り当て番号局（Internet Assigned NumbersAuthority）及びWHOISデーターベースで見つけることができる。

【0035】図2を、環境200を監視する分散されたサイトを図示するのを助けるのに使用する。典型的なメインコンテンツサーバサイト202は、例えば、定義されたリモートサーバ204，206，208，210及び212のような、分配されたサイトスイッチで稼働するVIPの稼働に対応する定義されたREAL SERVERのセットにアクセスする。各メインサイト202は、周期的なヘルスと、各定義されたリモートサーバのスループットチェックである。そして、各スイッチは、分配されたサイトスイッチにおけるVIPの稼働に対応する各々のその定義されたリモートREAL SERVERをテストする。各リモートサーバ204，206，208，210及び212に対する構成可能な反復ヘルスチェックを実行することによって、メインサイト202は、平均応答時間と、ハンド・オフに関する準備におけるコンテンツの可用性とを学習することができる。

【0036】これらのコンテンツヘルスチェックは、ヘルスチェックの全ての反復に関して、開始時間から終了時間まで、測定されるのが好ましい。サイト及びスイッチは、相互変換可能に使用されうる。サイトあたりのあるスイッチは、この例において仮定される。

【0037】図2において、分配されたサーバスイッチ202は、その好ましいハンド・オフサイトが、優先順に、定義されたリモートサーバ210，204，206，208であるように決定されうる。定義されたリモートサーバ210の900msec応答は、他のより遅い応答よりもより魅力的である。各リモートサーバ210，204，206，208の応答時間は、時間重み平均としてメインサイト202で記録される。この情報はまた、分配されたサイトステータスプロトコルを使用して他の全てのスイッチに各スイッチによって連絡される。互いのスイッチは、その定義されたリモートリアルサーバの各々に関して応答時間とスループットテストを行い、テストの開始からテストの終了までの合計のレスポンスインターバルを計算する。

【0038】例えば、HTTP，FTP，NNTP，DNS，SMTP及びPOP3のようなコンテンツヘルスチェックサポートを有するプロトコル及びアプリケーションに関して、コンテンツは、Adminによって定義された、例えばURL，ファイルネーム等のようなコンテンツコンフィギュレーションに基づいて反復的にアクセスされうる。コンテンツヘルスチェックでサポートされていないプロトコル及びアプリケーションに関して、又は、コンテンツコンフィギュレーションがまだ定義されていない場合において、TCO OPEN/CLOSE接続プロセスは、サーバロード平衡に関するほとんど同じ情報を作り出すように実行されうる。

【0039】図2では、分配されたサーバスイッチ106に対する4つの分配されたサイトのセットがある。ヘルス／スループットチェックは、分配されたサイトVIPに応答して各々定義されたリモートサーバに関して行われる。各サイトで対応するリモートリアルサーバを有する分配サーバスイッチ106で定義された5つのVIPがあるならば、分配されたサーバスイッチ106でスイッチは、（おのおの5つのリモートVIPを備える、4つの分配されたサイト）ヘルスチェックインターバルにわたって20のヘルス／スループットチェックをする必要がある。

【0040】リアルサーバヘルスは、リアルサーバに構成されたサービスに対する一連のTCP－SYN要求を介してテスト装置において監視される。これらの要求は、デフォルトによって数秒ごとに行われる。どんな反応の遅いサーバでも、サーバが「ダウン」を宣言するか、応答可能になるまで、反復的に要求を受信する。

【0041】別の考えは、ヘルスチェック中に個々のスイッチがリモートサーバに届かないならば、個々のスイッチがすべきであるということである。この状況が生じたとき、他のスイッチにもはや接続することができないスイッチは、(a)接続ハンド・オフに関してサーバスイッチがもはやふさわしくないと判断されるべきであり、ドメインネームサーバ応答又は「HTTP転送」に関する目標としてリモートサーバのVIPを使用して停止すべきであり、(b)サーバスイッチが応答可能でない他の全ての分配されたサイトに知らせるために、分配されたサイトステートプロトコル（DSSP）トリガー更新を送信すべきである。他の全てのサイトは、次いで、サーバスイッチが応答可能か、それに応じて作動するかどうか判断する。

【0042】分配された分配サーバステートプロトコル（DSSP）は、あるサイトから他の分配された分配サーバごとにステータス及びヘルス情報を連絡するのに使用される。プロトコルは、(a)これが通常及び周期的なUPDATEか、又は、これがEVENT通知であるか？、(b)VIPハンド・オフオーダされたリスト及び重み付けされた平均応答時間、(c)VIPあたりの利用可能な接続のような分配されたサーバ容量を残し、スイッチにおけるメモリリソース可用性を残すことを判断する能力である。

【0043】通常の周期的なリアルサーバヘルスチェックプロトコルが、サイトが応答可能であるか否かを判断するので、「keep-alive」又は「hello-are-you-there?」プロトコルとしてDSSPを使用する必要がない。

【0044】表1は、図2と同様な、サイトあたり単一のVIPを備えるサイトA－Fを備える仮定のネットワークにおけるシミュレーションされた応答時間を示す。時間は、各々のサイトの点から見たものである。本発明

の実施形態では、表1によって表されたものと同様の情報の表は、ＤＳＳＰを使用するサイトの間で通信される。各受信サイトは、後で使用するためにＶＩＰハンド・オフオーダーされたリストを生成するためにスループット数の比較を行う。各サイトＡ－Ｆでの各スイッチは、テストされる分配されたサーバがどんなヘルスチェックにも応答しないならば、テストサイトの透視から「ダウン」していると判断できる状況を除いて、同じハンド・オフテーブルを計算する。

【００４５】表1

――――――――――テストをするサイト――――――――

テストされたサイト	A	B	C	D	E	F
A	*	3155	1073	3439	113	641
B	2925	*	1314	378	813	1827
C	1364	207	*	3869	995	3883
D	197	2490	1997	*	1190	339
E	3702	1106	1743	2344	*	468
F	1759	1409	683	2235	419	*

（ミリ秒における平均遅延時間）

これらの測定によりサイトＡでは、サイトＤが高いスループットであるように見える。サイトＢは、高いスループットを有するようにサイトＣが見え、サイトＣ及びサイトＥはサイトＦが高いスループットを有するように判断する。

【００４６】表2は、表1で測定された、各サイトのオーダされたハンド・オフ・プリファレンスの結果である。この情報が、サイトの間で交換されるとき、各サイトは、各サイトが最初のプリファレンス、第2のプリファレンスなどとなるのにどれだけの時間がかかるか計算する。

【００４７】表2

――――――――サイトプリファレンス選択――――――――

オーダ	A	B	C	D	E	F
1	D	C	F	B	A	D
2	C	E	A	F	F	E
3	F	F	B	E	B	A
4	B	D	E	A	C	B
5	E	A	D	C	D	C

【００４８】表2において、サイトＡはある例では第1のプリファレンスである。サイトＢは、ある例では第1のプリファレンスである。サイトＣはある例では第1のプリファレンスである。サイトＤはある例では第1のプリファレンスである。サイトＥはある例では第1のプリファレンスである。サイトＥは決して現れない。そして、サイトＦはある例では第1のプリファレンスである。第2行では、Ａ＝１，Ｂ＝０，Ｃ＝１，Ｄ＝０，Ｅ＝２，Ｆ＝２を生成する。

【００４９】表3：静的重みテーブル

ＤＮＳ／ＨＴＴＰ転送ハンド・オフ重み（Traff Dist付）

サイト	総重み	ﾄﾗﾌｨｯｸ disti	オーダ	所定の重み
A	7	１７％	重み１	4
B	6	１４％	重み２	2
C	6	１４％	重み３	1
D	8	１９％	重み４	0
E	5	１２％	重み５	0
F	１０	２４％	重み６	0

【００５０】表3の所定の重みの列を参照すると、各最初の位置の外観は、第3の位置の外観の重みの4倍受信

するのが好ましい。各第２の位置の外観は、第３の位置の外観の重みの２倍受信する。第４から第６の位置の外観は、重みなく受信する。従って、本発明のアルゴリズムの実施形態は、表４に示したように構成されうる。

【００５１】表４

サイトＡの「総重み」＝(1*4)+(1*2)+(1*1)=7;
サイトＢの「総重み」＝(1*4)+(0*2)+(2*1)=6;
サイトＣの「総重み」＝(1*4)+(1*2)+(0*1)=6;
サイトＤの「総重み」＝(2*4)+(0*2)+(2*1)=10;
サイトＥの「総重み」＝(0*4)+(2*2)+(1*1)=5;及び
サイトＦの「総重み」＝(1*4)+(2*2)+(2*1)=10

【００５２】かかる方法を使用する際に種々の利点がある。最も良いサイトは一般的に、他のサイトよりもより多く通信を受信するが、より多くの通信は受信しない。発生するいかなるハンド・オフもトップ２乃至３のサイトにわたる平均であるのが好ましく、静的ハンド・オフ重みを調整することによって調整可能である。全ての他のサイトによって不十分に実行されるようなサイトは、少ししか受信せず、又は全くハンド・オフしない。ＷＡＮリンク、サーバなどを含む全てのサイトが上手く実行するならば、各サイトが時間にわたって等しいトラフィックの分配を受信しているようなものである。

【００５３】テーブルIIIのような計算されたハンドオフテーブルは、もっぱら、ＤＮＳ応答の順序付けや「ＨＴＴＰリダイレクト（転送）」プリファレンス（優先）に使用される。「ＨＴＴＰリダイレクト」が呼び出されないときにＴＣＰ接続リクエストがＶＩＰへ送られてきた場合、それは使用されない。

【００５４】モニタリング・ハンドオフ交換プロセスに３つのサイト若しくはそれより少ないサイトしか含まれていないときは、ハンドオフの決定におけるグラニュラリティの貧弱さが問題になることがある。このような場合、最も極端な場合を除けば、「最良」対「最悪」サイトを正確に決定するに十分なスループットデータサンプルは存在しない。このような状況におけるこの問題を緩和するため、スイッチ内の制御や同調可能パラメータが含まれていなければならない。多くの接続を受け入れることができるサイトは、多くの接続を受け取るために、あるテンデンシー（性癖）を有する。

【００５５】ＤＳＳＰトリガ型更新は、通常の更新が有する全ての情報を含んでいるのが好ましいが、これらのトリガ型更新は、あるスイッチが遠隔サーバともはや通信をすることができないときに、或いは、全てのサーバがそれらの各ＭａｘＣｏｎｎｓに存在し、リアルサーバをＶＩＰのために利用することができない等のように、あるスイッチがローカルリソースコンストレイント（局部資源制約）を被ったときに、そのスイッチから他の全てのスイッチへ直ちに送信される。

【００５６】ＤＳＳＰ更新の例を示すため、サイトＡは５つのピア（同一層内の）サイトＢ～Ｆを有する。各サイトＡ～Ｆは２つのＶＩＰを実行し、また、各サイトと互いにピアされている（同一の層内とされている）。セッションハンドオフ分散型サーバの決定のため、各サイトのスイッチは、各遠隔ＶＩＰ／局部ＶＩＰ結合についての各マッチング（整合）ドメインネームのために、順序付けされたハンドオフテーブルを計算する。各スイッチは、“www. Alteon. com”を表示するＶＩＰを伝達し、エントリが、各ＶＩＰの試験応答性に基づいて「計算されたハンドオフテーブル」に現れる。“www. Alteon. com”のような所定のドメインネームのため、順序付けされたハンドオフテーブルが各スイッチによって構築されるのが好ましい。ハンドオフテーブルはその後、テーブルを構築するためのドメインネームに関するドメインネームサーバリクエストがスイッチによって受け取られたときに、調査される。各スイッチは、本明細書のテーブルに示されているように、計算された重み値に基づいて遠隔リアルサーバの重みを動的に更新する。“www. Alteon. com”についてのドメインネームサーバリクエストをいずれかのスイッチが受け取ったとき、そのスイッチは、現在の重みに基づいて「次に好適の」遠隔サーバに対応するＩＰアドレスを用いて応答する。分散型サーバＦに対応するＶＩＰは、大体、リクエストの２５％を受け取る。換言すれば、いずれかのサーバが受け取っている時間の２５％はドメインネームサーバリクエストである、ということであり、そのスイッチは分散型サーバのＶＩＰアドレスを用いて応答する。

テーブルＶ：順序付けされたハンドオフ重みテーブル
ＤＮＳ／ＨＴＴＰリダイレクトハンドオフ重み（トラフィック識別を有する）

サイト	重みの総計	トラフィック識別	順序
A	11	26%	重み－1
B	12	29%	重み－2
C	0	0%	重み－3
D	6	14%	重み－4
E	10	24%	重み－5
F	3	7%	重み－6

【００５７】順序付けされたハンドオフテーブル、即ち、テーブルVでは、サイトCは0の重みを有する。このＶＩＰは、このＶＩＰに送信されるいずれのハンドオフリクエストをも有してはならない。この例では、サイトA、BおよびEがハンドオフの大部分を受け取る。

【００５８】セションハンドオフ実行のため、スイッチは、自身がホストになっている（hosting）ドメインネームについてのドメインネームサーバリクエストを受け取ったときに、ハンドオフ重みやアベイラビリティ（可用性）等に基づいて、それらのドメインをロードバランス（負荷平衡化）しているスイッチの適当なＩＰアドレスを用いて応答する。一般に、近くのサーバがダウンしているかまたはオーバーロードしていない場合には、ある領域内のユーザがその領域内の若しくはその領域付近のサーバと関連付けられているときに最良であることが好ましい。例えば、“www. Alteon. com”に関するホストコンテント（内容）がインストールされた５つのサイトが全世界、即ち、サンジョーズ（西アメリカ）、アトランタ（東アメリカ）、エクアドル（南アメリカ）、パリ（フランス）、東京（日本）に存在しているものとする。ヨーロッパのユーザは、パリのサイトによってサービスを受けるのが好ましく、チリのユーザは、エクアドルのサイトによってサービスを受けるのが好ましい等である。日本のユーザがコンテントのためにはるばるアトランタのサイトへやって来たと仮定すると、多くの他のユーザがそこから利益を受けることができるような帯域幅を浪費させ、また、このようなサービスは、直接的には、日本のユーザにも不必要な応答遅延を生じさせるものとなる。

【００５９】それゆえ、どんなセッションハンド・オフを実行する前にもユーザ要求の地理的ソースを最終的に決定するために重み付けをすることがスイッチに関して重要である。スイッチが、構成されたドメインに関するドメインネームサーバ要求を受信するとき、要求、及びＩＡＮＡから種々の地域の登記に発行されたＩＰアドレスブロックを備えるそれに全体的に関するソースＩＰアドレスを調査すべきである。表４は、種々の地域の登記に関するアドレスブロック割り当てのいくつかを示し、それらはそれぞれ地理的なドメインである。

表VI

RIPE NCC－ヨーロッパ 97年4月	063/8
ARIN 97年4月	064-095/8
RIPE NCC－ヨーロッパ 93年5月	194/8
RIPE NCC－ヨーロッパ 93年5月	195/8
RIPE NCC－ヨーロッパ 93年5月	196/8
ARIN－北アメリカ 93年5月	199/8
ARIN－北アメリカ 93年5月	200/8
ARIN－中央及び南アメリカ 93年5月	201/8
ARIN－中央及び南アメリカ 93年5月	202/8
APNIC－環太平洋地域 93年5月	203/8
APNIC－環太平洋地域 93年5月	204/8
ARIN－北アメリカ 94年3月	205/8
ARIN－北アメリカ 94年3月	206/8
ARIN－北アメリカ 95年4月	207/8
ARIN－北アメリカ 95年11月	208/8
ARIN－北アメリカ 96年4月	209/8
ARIN－北アメリカ 96年6月	210/8
APNIC－環太平洋地域 96年6月	211/8
APNIC－環太平洋地域 96年6月	212/8
RIPE NCC－ヨーロッパ 97年10月	213/8
ARIN－北アメリカ 98年4月	217/8

【0060】表VIのエクステンションは、本発明の各スイッチの実施例がアクセスできるデーターベース形式で提供されるのが好ましい。ソースネットワークは、124－ビットIPサブネット深度に変換されるのが好ましい。使用されるデーターベースは、IANA「WHOIS」データベースから得られるのが好ましい。スイッチにおいてこのような表の情報を使用することによって、ドメインネームサーバーが、ドメインネームサーバーレスポンダーは、要求に関して大よその地理的な決定をすることが可能となる。もし、ドメインネームサーバー要求が211.123.11.20であるならば、要求しているホストは環太平洋地域ののいずれかに位置し、203、204、211、212のいずれかで始まるサイトに向けられるべきである。もし、分散サイトVIPのいずれかが地理的に異なったネットワークにあるならば、スイッチは、全ドメインネームサーバーが応答する間、この表の情報を使用するのが好ましい。

【0061】ピアハンドーオフプロセスでは、スイッチは、特定のVIPドメインネームのドメインネームサーバールックアップ要求を受信する。スイッチは、ドメインネームサーバー要求のソースIPアドレスを調べ、ユーザーのIPアドレスを調べ、そのユーザーに地理的に近いサーバーサイトがあるかどうかを決定する。スイッチは、ドメインに対応する順序づけられたハンドーオフ表を調べる。スイッチは、（a）ドメインネームサーバー要求ソースと比較されたリモートサーバー位置、（b）リモートサーバー重み、（c）先のハンドーオフを体験したリモートサーバーに基づいて、好ましく、次のリモートサーバー（もしくはそれ自体のVIP）を選ぶ。そして、スイッチは、ドメインネームサーバー応答を順序付けられたリストのIPアドレスと共にクライアントドメインネームサーバーに返す。

【0062】スイッチが、スイッチVIPへの「TCP SYN」を受信する時、それは、パケットを受け取るか、もしくは、もし、ローカルVIPが過負荷ならば、パケットを拒絶する。もし、拒絶ならば、スイッチは、このドメインの順序付けられたハンドーオフ表を調べ、（a）ドメインネームサーバー要求ソースと比較されたリモートサーバー位置、（b）各リモートサーバーの重み、（c）先のハンドーオフで識別されたリモートサーバーに基づいて、好ましく、次のリモートサーバーもしくはそれ自体のVIPを選ぶ。スイッチは、負荷及び他のサイトの使用可能性に応じて、「HTTPリダイレクト」をクライアントに返すか、要求を捨てる。

【0063】スイッチが、ドメインネームサーバー応答を出す時、それは、設定可能なドメインネームサーバー

TTL値でそのようにし、ダウンストリームドメインネームサーバーが、長時間、サーバースイッチのIPアドレスをキャッシュしないことを確実にする。

【0064】分散負荷バランシングパラメーターに関し、各スイッチはスイッチーワイド分散SLBーパラメータで構成されて、その分散サイトを識別するのが好ましい。例えば、全ての他のスイッチの管理IPーアドレスのリストによって構成される。

【0065】様々な調整可能なパラメータが、本発明の実施例に含まれるのが好ましい。スイッチあたり8つの設定可能な分散サイトを有する分散サイトが、リモートスイッチのIPーアドレスで構成される。これらのサイトの各々は、リモートリアルサーバー（VIPの）が存在する潜在的なハンドーオフサイトとしてスイッチによって認識され得る。分散ーサーバー状態プロトコルインターバルは、スイッチが、どのくらい頻繁に定期的なDSSPアップデートを通知するかを表す。1分のデフォルトで1－120分の範囲が選ばれ、個々のサイト毎に変えられる。ドメインネームサーバーTTLは、ドメインネームサーバー要求に応答する時に使用されるべきTTLー値を表す。1分のデフォルトで0－225分の範囲が選ばれる。分散SLBオン／オフ制御に関し、「HTTPリダイレクト」オプションを使用でき、デフォルトを「オン」として「オン／オフ」にセットできる。また、「UseDNSRespond」オプションを使用でき、デフォルトを「オン」として「オン／オフ」にセットできる。順序づけられたハンドーオフ重み（1－16にインデックス付けされた）は1－64の値を有することができ、順序づけられたハンドーオフリストを計算する間考慮される。

【0066】各ハンドーオフ重みインデックス（1、2、3、・・・16）は、最良のパフォーミングサイト対最悪のパフォーミングサイトに相当する。各インデックスは、順序付けられたハンドーオフリストにおけるサーバースイッチの相対位置によって乗算されるのが好ましい統計的に設計された重みを有することができる。もし、順序付けられたハンドーオフ重み（OHW）インデックスー1が4にセットされるならば、最良のパフォーミングサイトは、1の重みを有するサイトの4倍の接続を受信する。代表的な構成を、OHW－1を「6」に、OHW－2を「4」に、OHW－3を「2」に、そして他の全てを「1」にといったようにセットしても良い。これは、第1、第2、第3の最良のパフォーミングサイトを、サーバースイッチの残りのものと比較して6倍、4倍、2倍の多くのハンドーオフを受信するように導く。

【0067】図3は、本発明の分散サーバーウェッブバランス方法の実施例のフローチャートを表し、ここでは、共通参照番号300で参照される。この方法300はステップ302で始まり、そのステップでは、DNSールックアップのユーザー要求が受信される。このような要求は、特定のウェッブーベースの内容及びサービスで応答するIPーアドレス番号を要求する。ステップ304は、DNSールックアップ問合せに含まれるユーザーIPーアドレスを調べることによって、ユーザーの地理的なドメインが何であるかを決定する。ステップ306は、ユーザーの地理的なエリア中もしくはユーザーの地理的なエリア近くの使用可能なネットワークサイト及びスイッチを調べる。ステップ308は、以降、ユーザーと対応するジョブを与えられるべき「最良の」仮想IPーサーバー（VIP）を計算する。何が「最良」を生ぜしめるかは、何の目標が検討されるかによって変わる。「最良」は、ユーザー、ウェッブサイト、バックボーンオペレータ、インターネットサービスプロバイダー（ISP）、コスト等々の観点からの最良のシステム全体の性能であり得る。バックグランドプロセス310は、全てのVIPの稼働状態及び性能を継続的に監視する。ステップ312は、ユーザーのDNSールックアップ要求に、ユーザーにサービスするための「最良の」VIPのIPーアドレスで応答する。

【0068】本発明を好ましい実施例を参照してここに説明したけれども、本発明の精神及び範囲から逸脱することなく、他の応用をここに説明したものの代わりにしても良いことは当業者は容易に理解されよう。従って、本発明は、以下に含まれる請求の範囲によってのみ限定されるべきである。

【図面の簡単な説明】

【図1】図1は、本発明の分配されたサーバのロードバランシングシステム実施形態のブロックダイアグラムである。

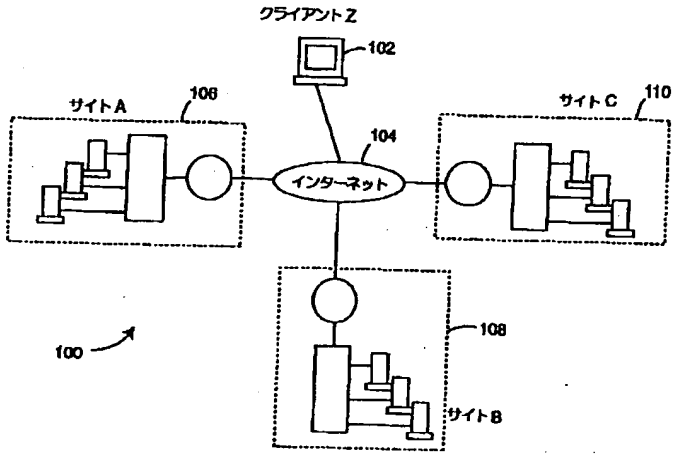
【図2】図2は、サイトAがウェブページのアクセスのためにクライアントリクエストを冗長して支持できる他のいくつかのサイトについて得られる情報を示すダイアグラムである。

【図3】図3は、本発明の分配されたサーバのロードバランシング方法実施形態のフローチャートである。

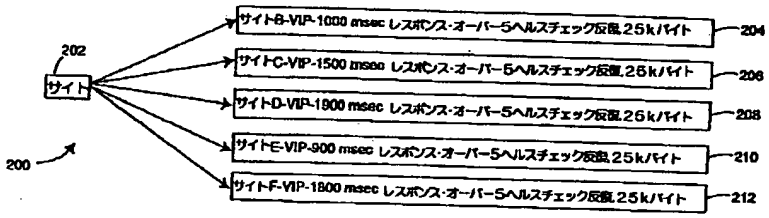
【符号の説明】

102　クライアント
104　インターネット
202　サイト

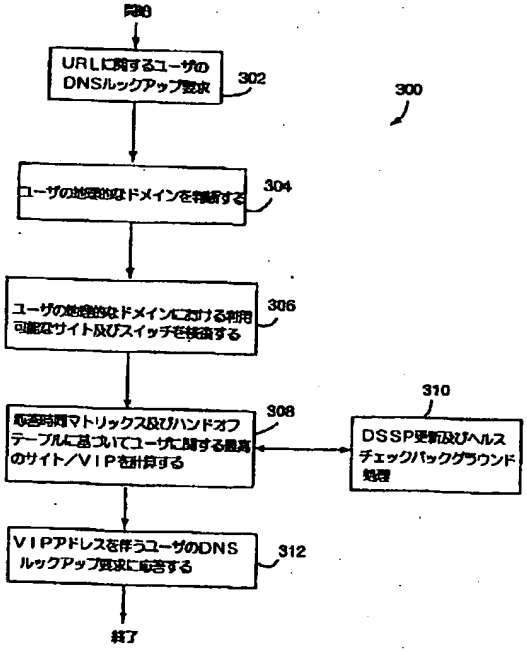
【図1】


クライアントZ
102
サイトA
106
サイトC
110
104
インターネット
108
100
サイトB

【図2】


202
サイト
200
サイトB-VIP-1000 msec レスポンス-オーバー5ヘルスチェック反復 25kバイト
204
サイトC-VIP-1500 msec レスポンス-オーバー5ヘルスチェック反復 25kバイト
206
サイトD-VIP-1900 msec レスポンス-オーバー5ヘルスチェック反復 25kバイト
208
サイトE-VIP-900 msec レスポンス-オーバー5ヘルスチェック反復 25kバイト
210
サイトF-VIP-1800 msec レスポンス-オーバー5ヘルスチェック反復 25kバイト
212

【図3】


開始
URLに関するユーザのDNSルックアップ要求
302
300
ユーザの地理的なドメインを特定する
304
ユーザの地理的なドメインにおける利用可能なサイト及びスイッチを検索する
306
310
最適時間マトリックス及びハンドオフテーブルに基づいてユーザに関する最高のサイト/VIPを計算する
308
DSSP更新及びヘルスチェックバックグラウンド処理
VIPアドレスを伴うユーザのDNSルックアップ要求に応答する
312
終了

フロントページの続き

(72)発明者　シリッシュ　サテイ
アメリカ合衆国　カリフォルニア州
95118　サン　ホセ　ボールデラ　ドライヴ　1533

【外国語明細書】

DISTRIBUTED LOAD-BALANCING INTERNET SERVERS

BACKGROUND OF THE PRESENT INVENTION

TECHNICAL FIELD

The present invention relates generally to computer data network equipment and methods, and more particularly to balancing the loading amongst distributed network servers by controlling the conversion of domain names to IP-addresses in domain name server equipment. The invention selects the load distribution criteria based on a unique algorithm.

DESCRIPTION OF THE PRIOR ART

The world wide web (WWW), and especially the Internet, are quickly becoming the principle way businesses sell products and communicate with customers and suppliers. Some now call the Internet a "mission-critical business delivery infrastructure." As a consequence, Internet servers and so-called "Intranet" servers are worked harder than ever before. The number of clients servers now must support has increased dramatically. Intranet servers must now be able to service hundreds of simultaneous client requests, while their external-counterpart Internet servers must be able to support tens of thousands of simultaneous client connections.

Clients demand and expect rapid response and a 7-day and week, 24-hours a day ("7 x 24") availability. Mission-critical web-computing infrastructures must be able to dynamically scale server capacity to match aggregate client demand and still ensure continuous service availability. One way to do just that has been to run each application on several servers, and then continually balance the client loading on the various servers, e.g., "server load balancing."

Server load balancers use information in the Layer 3 and Layer 4 packet headers to identify and manage application-layer sessions. For example, TCP or UDP port numbers, the

SYN/FIN bits that mark the start and end of TCP application sessions and IP source and destination addresses.

Traditional server load balancers are PC-based software products with limited performance and connectivity. The rapid growth in traffic volume and server population is giving rise to a new generation of switch-integrated server load balancers that offer many orders of magnitude improvements in performance, connectivity, resiliency and economy

A new generation of switch-based server load balancers consolidates multiple web infrastructure functions and load balancing application servers with multi-layer switching, e.g., redirection traffic to caches, load balancing traffic to multiple firewalls, packet filtering and bandwidth management.

Alteon WebSystems coined the term "Server Switch" to represent this new class of device that front-ends server farms and provide server-related traffic management in al mission critical internet/intranet infrastructures. Server Switches dynamically distribute application load across a group of servers running a common application (or set of applications) while making the group appear as one server to the network. A number of web servers with access to the same content can be logically combined into an HTTP hunt group, which is a group of servers that supports a common application or set of applications. The hunt group provides a "virtual" HTTP service to clients. Clients are not aware that there are a number of real servers participating in providing this service. The clients access the service using a virtual service address that resides in a server switch that front-ends the real servers. As connection requests arrive for the virtual service, the server switch passes these requests on to one of the real servers in the hunt group based upon knowledge of the servers' availability, load handling capability, and present load.

In this way, multiple servers can be used to achieve the total amount of application processing capacity demanded by the users of the system. Each new server adds its capacity to the pool of processing power available for the application.

Equally important, as servers go out of service due either to failure or maintenance operations, the remaining healthy servers pick up the load with little or no perceived impact to users. To achieve this, the server switch must continuously monitor the health of al servers and each application to which it distributes client load. The server switches must also support hot-standby configurations for complete systems redundancy.

A key part of server load balancing is session management. Once a session request is assigned to a real server, the server switch must recognize all successive packets associated with that session. These packets are processed and forwarded appropriately to make sure that the client continues to be associated with the same physical server for the duration of each session.

Server switches also monitor the completion of sessions at which time the binding of the connection to the physical server can be removed. This ensures that the next time a client connects, he is preferably connected to the most available server at the time, providing the best possible service to each client. Special mechanisms can be invoked by the administrator if the application requires successive connections to be forwarded to the same physical server, such as with FTP control and data connections, SSL (Secure Sockets Layer), and persistent HTTP used for multi-page forms and search engines.

Environments that benefit from server load balancing include web hosting services, on-line service providers and corporate data centers with high availability requirements. In theory, server load balancing can be used to support any TCP-based or UDP-based application where common content is available across a group of servers. In practice, servers supporting Internet/Intranet applications, such as web servers, FTP servers, domain name server servers and RADIUS servers is preferably the first to take advantage of server load balancing to support the high growth and unpredictable volume of web-oriented traffic.

The majority of web pages contain read-only information. This makes web-hosting environments ideal for server load balancing. Web hosts and on-line service providers typically deploy multiple HTTP, FTP and other application servers today, with load distributed across them statically, or more commonly, via round-robin domain name server. Both methods are undesirable because they are not fault-tolerant and require a high degree of administration. Server load balancing enables transparent use of multiple servers with built-in high availability support.

Many clustering systems today provide superior failover capabilities but offer no load-balancing support. Some systems also limit the number of servers that can participate in a cluster. These constraints impact the scalability of the clustering solutions. Server load balancing enables flexible coupling of servers into load-sharing hunt groups. It also improves server utilization efficiency by enabling redundant servers to share load.

More often than not, server environments today are multi-vendor and multi-OS. Popular clustering solutions today are limited to servers from a single vendor or servers running a single operating system. Server load balancing on a server switch enables heterogeneous servers supporting TCP and UDP applications to be loosely coupled in a load-sharing cluster, maximizing server investment returns.

SUMMARY OF THE PRESENT INVENTION

An actual Internet web-site that serves the web-pages to a client in response to a URL domain name is automatically and transparently selected from a list of many distributed sites each having identical data storage. In a peer hand-off process, a switch receives domain name server lookup request for a particular domain name. The switch examines the source IP-address for the domain name server request, examines the user's IP-address, and determines if there is server site that is geographically close to that user. The switch examines an ordered hand-off table corresponding to the domain. The switch chooses a next remote server (or one of its own virtual Internet protocol addresses) based on, (a) the remote server location compared to domain name server request source, (b) the remote servers' weights, and (c) the remote server that experienced the previous hand-off. The switch then sends the domain name server response back to client domain name server with the IP-addresses in an ordered list.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a block diagram of a distributed-server load-balancing system embodiment of the present invention;

Fig. 2 is a diagram illustrating the information a site-A can obtain about several other sites that could redundantly support client requests for web-page accesses; and

Fig. 3 is a flowchart of a distributed-server load-balancing method embodiment of the present invention.

DETAILED DESCRIPTION OF THE PRESENT INVENTION

Fig. 1 represents a distributed-server load-balancing system embodiment of the present invention, and is referred to herein by the general reference numeral 100. The distributed-server load-balancing system 100 allows web-based content and services to be redundantly delivered to many clients, represented by a client "Z" 102, from many independent web-server sites over Internet 104. When a client 102 loads a web-browser program and enters a uniform resource location (URL), e.g., "www.alteon.com/products/index.html".

While IP-addresses used on the Internet 104 are 32-bits in length, most users do not memorize the numeric addresses of the hosts to which they attach. Instead, people are more comfortable with host names. Most IP hosts, then, have both a numeric IP-address and a name. While this is convenient for people, however, the name must be translated back to a numeric address for routing purposes. Internet hosts use a hierarchical naming structure comprising a top-level domain (TLD), domain and subdomain (optional), and host name. The IP-address space, and all TCP/IP-related numbers, is assigned and maintained by the Internet Assigned Numbers Authority (IANA). Domain names are assigned by the TLD naming authority; until April 1998, the Internet Network Information Center (InterNIC) had overall authority of these names, with NICs around the world handling non-U.S. domains. The InterNIC was also responsible for the overall coordination and management of the domain name System (DNS), the distributed database that reconciles host names and IP-addresses on the Internet.

In client-Z 102, a domain name server "getByHostname" query is actually issued to a local domain name server, asking for the numeric Internet Protocol address (IP-address) that has been registered for use with "www.alteon.com". Each local domain name server checks to see if it already knows the IP-addresses for the hosts that service particular domain name and host. It could know this by having previously needing this information and storing the answer it discovered in a local private cache memory. If the local domain name server does not know the hostname IP-address for a requested URL domain name, it will perform an iterative query to a domain name server higher in the DNS hierarchy. Such domain name server query will either be answered by a higher level domain name server, or the request will ultimately bubble up to one of a distributed-server network switch sites 106, 108, or 110.

IP-addresses are hierarchical for routing purposes and are subdivided into two subfields. The Network Identifier (NET_ID) subfield identifies the TCP/IP subnetwork connected to the Internet. The NET_ID is used for high-level routing between networks, much the same way as the country code, city code, or area code is used in the telephone network. The Host Identifier (HOST_ID) subfield indicates the specific host within a subnetwork.

Most IP hosts usually have both a numeric IP-address and a name. The name is provided as a convenience for people, however such name must be translated back to a numeric address for routing purposes. Internet hosts use a hierarchical naming structure comprising a top-level domain (TLD), domain and subdomain (optional), and host name. The distributed-server network switches 106, 108, and 110 are organized as distributed sites, where each acts as an Authoritative Name Server for a sub-domain, e.g., "www.alteon.com". Each such distributed site is capable of responding to a domain name server query with the IP-address identities that correspond to "www.alteon.com".

The TCP/IP protocol suite comprises two protocols that correspond roughly to the OSI Transport and Session Layers. These protocols are called the Transmission Control Protocol and the User Datagram Protocol (UDP). Individual applications are referred to by a port identifier in TCP/UDP messages. The port identifier and IP-address together form a socket. Well-known port numbers on the server side of a connection include port -20 (FTP data transfer), port -21 (FTP control), port -23 (Telnet), port -25 (SMTP), port -43 (whois), port -70 (Gopher), port -79 (finger), and port -80 (HTTP).

For illustration purposes, assume that the distributed-server switch 108 receives a domain name server query that originated with client 102. In embodiments of the present invention, the distributed-server switch 108 will return a set of IP-addresses that represent a virtual-IP (VIP). For example, the distributed-server switch 108 could respond to the URL query with a set of IP-addresses including "192.168.13.20", "162.113.25.28", and "172.176.110.10", any one of which could satisfy web-based content and service demands associated with the single URL. Each of these several IP-addresses exists at a geographically diverse server, e.g., as represented by distributed server switches 106 and 110. The client 102 will receive such response

via its local domain name server. The client 102 is then able to use those IP-addresses and open a TCP Port 80 connection to "192.168.13.20" which is, for example, a VIP-address actually running at distributed-server switch 106. The client 102 does not know this is only a VIP, and can ignore a real IP-address of "192.168.13.10" that exists at switch 106. Thereafter, the traffic generated by client 102 with the "www.alteon.com" website is handled by the distributed-server switch 106 and off-loaded from the other possible switches 108 and 110.

The VIP's set up for each switch 106, 108, and 110 must each enable client access to the same content and applications, so that a request to any one will result in the same data being given to the client 102. A policy therefore needs to be established that distributes the available resources to the users needing service. The factors to consider in such policy include the health of the individual distributed-server VIP's involved, the basic Internet assigned numbers authority (IANA) registered location of the client and server(s), and a list of the available servers according to currently measured response times and throughputs. Those servers that are the healthiest, more closely located, and showing good response times and throughputs should have more of the traffic directed to them. This is done by responding with their corresponding VIP's more often.

The DNS is a conventional distributed database of host name and IP-address information for every domain on the Internet. There is a single authoritative name server for every domain. About a dozen root servers have a list of all of these authoritative name servers. When a request is made by a host to the DNS, the request goes to a local name server. If there is insufficient information at the local name server, a request is made to the root to find the authoritative name server, and the information request is forwarded to that name server. Name servers contain the following types of information:

A-record: An address record maps a hostname to an IP-address.
PTR-record: A pointer record maps an IP-address to a hostname.
NS-record: A name server record lists the authoritative name server(s) for a given domain.
MX-record: A mail exchange record lists the mail servers for a given domain.

If the server switch 106, 108, or 110, that client 102 has been pointed to suddenly experiences a failure or is overloaded, it will issue an ""HTTP redirect"". The client 102 is thus commanded to go to a different server switch 106, 108, or 110. The ""HTTP redirect"" will occur when an "HTTP Request" arrives at a VIP that is at maximum connections ("MaxConns") or no longer has any healthy real servers.

The distributed-server load-balancing system 100 of Fig. 1 uses a domain name server to respond to DNS-requests for VIP sites. The "www.alteon.com" example represents several VIP's scattered through the United States with access to the same content for the Alteon Web distributed-server. When the switch receives a domain name server Name Request to resolve "www.alteon.com", associated with a VIP, it will respond with an appropriate domain name server response that matches the "best site" to respond to the subsequent content requests. Such best site, for example, represents the one that imposes minimum delays on the greatesat numbers of users. Other criteria are possible, such as defining the best site to respond as the one that is the least costly.

Site health and throughput measurement is obtained during "L4 health-checking" (with content verification as an option) with all the other peer remote sites 106, 108, and 110. Such is used to determine the status of the application availability and also the throughput performance of each site.

A distributed SLB state protocol is used that is capable of exchanging health, load and throughput information between sites either periodically, or when triggered by a predefined event. An Internet topology awareness is preferably included in embodiments of the present invention.

For Internet topology awareness, the particular switch used for DNS/HTTP hand-offs will examine the Source_IP for the request, and will respond with a "best" server based on the IANA allocated IP-address space throughout the world. Other hand-off criteria is also included. An external "subscribers database" may be required to provide the necessary amount of detail that describes where registered user networks are located. This information can be found at the Internet Assigned Numbers Authority and the WHOIS database.

Fig. 2 is used to help illustrate distributed site monitoring environment 200. A typical main content server site 202 has access to a set of defined REAL SERVER's which correspond to VIP's running in distributed site switches, e.g., defined remote servers 204, 206, 208, 210, and 212. Each main site 202 does a periodic health and throughput check of each defined remote server. And each switch tests each of its defined remote REAL SERVER's which correspond to VIP's running in distributed-site switches. By executing a configurable iterative health-check to each remote server 204, 206, 208, 210, and 212, a main site 202 can learn the average response times and content availability in preparation for a hand-off. These content health-checks are preferably measured from start-time, to end-time, for all iterations of the health-check. Site and switch can be used interchangeably. One switch per site is assumed in this example.

In Fig. 2, the distributed-server switch 202 could determine that its preferred hand-off sites are defined remote servers 210, 204, 206, 208, in order of priority. The 900 msec response of defined remote server 210 is more attractive than the slower responses of the others. The response times of each remote server 210, 204, 206, 208 are recorded at main site 202 as a time-weighted average. This information is also communicated by each switch to all other switches using distributed-site status protocol. Each other switch does response time and throughput tests for each of its defined remote real servers, and computes total start-of-test to end-of-test response interval.

For applications and protocols that have content health-checking support, e.g., HTTP, FTP, NNTP, DNS, SMTP, and POP3, the content can be iteratively accessed based on the content configuration, e.g., URL, filename, etc., as defined by the Admin. For applications and protocols not supported with content health-checking, or in cases where the content configuration has not yet been defined, a TCP OPEN/CLOSE connection processes can be executed to produce nearly the same information for the server load balancing.

In Fig. 2, there are a set of four distributed sites to distributed-server switch 106. A health/throughput check is done for each defined remote server corresponding to a distributed site VIP. If there are five VIP's defined at distributed-server switch 106 which have corresponding Remote REAL SERVER's at each site, the switch at

distributed-server switch 106 will have to do 20 Health/Throughput checks over the health-check interval (four distributed sites, with five Remote VIP's apiece).

Real server health was monitored in test equipment through a series of TCP-SYN requests to the services that are configured on the real servers. These requests took place every few seconds by default. Any unresponsive servers would receive iterative requests until the server was declared "down" or became responsive.

Another consideration is what an individual switch should do if it cannot reach a remote server during health-checks. When this situation occurs, the switch that no longer can communicate to another switch should (a) no longer consider the server switch eligible for connection hand-offs, and stop using the remote server's VIP as a target for domain name server responses or "HTTP redirects"; and (b) send out a distributed site state protocol (DSSP) triggered update to inform all other distributed sites that the server switch is not responsive. All other sites may then determine if the server switch is responsive and act accordingly.

The Distributed distributed-server State Protocol (DSSP) is used to communicate Status and Health information from one site, to every other Distributed distributed-server. The Protocol is capable of determining (a) is this a normal and periodic UPDATE or is this an EVENT notification?, (b) a VIP hand-off ordered list and weighted average response times, (c) any remaining distributed-server capacity such as connections available per VIP and remaining memory resources available in the switch.

It is not necessary to use DSSP as a "keep-alive" or "hello-are-you-there?" protocol, because the normal periodic Real server health-checking protocol will determine whether a site is responsive or not.

Table I represents the simulated response times in a hypothetical network with sites A-F with a single VIP per site, similar to that of Fig. 2. The times are with respect to each site's point of view. In embodiments of the present invention, tables of information, like that represented by Table I, are communicated between sites using DSSP. Each recipient site does comparisons of throughput numbers to create a VIP hand-off ordered list for use later. Each switch at each site A-F calculates the same

hand-off table, with the exception that if a tested distributed-server did not respond to any health-checks, it is considered as being "down" from the testing site's perspective.

TABLE I

site tested \ site doing the test	A	B	C	D	E	F
A	*	3155	1073	3439	113	641
B	2925	*	1314	378	813	1827
C	1364	207	*	3869	995	3883
D	197	2490	1997	*	1190	339
E	3702	1106	1743	2344	*	468
F	1759	1409	683	2235	419	*

(average delay time in milliseconds)

It would appear to site-A with these measurements that site-D is high throughput. Site-B sees site-C as having high throughput, and site-C and site-E will determine site-F has high throughput.

Table II is the result of what each site's ordered hand-off preferences would be, given the measurements in Table I. When this information is exchanged between sites, each switch calculates how many times each site was first preference, second preference, etc.

TABLE II

order \ site preference choices	A	B	C	D	E	F
1	D	C	F	B	A	D
2	C	E	A	F	F	E
3	F	F	B	E	B	A
4	B	D	E	A	C	B
5	E	A	D	C	D	C

In Table II, site-A was first preference in one instance. Site-B was first preference in one instance. Site-C was first preference in one instance. Site-D was first preference in two instances. Site-E was first preference in one instance. Site-E never appeared. And, site-F was first preference in one instance. The second row produces A=1, B=0, C=1, D=0, E=2, F=2.

TABLE III: Static Weight Table

DNS/HTTP Redir Hand-off Weights (with Traff Dist)

site	total weight	traffic dist	order	given weight
A	7	17%	weight-1	4
B	6	14%	weight-2	2
C	6	14%	weight-3	1
D	8	19%	weight-4	0
E	5	12%	weight-5	0
F	10	24%	weight-6	0

Looking at the given weight column in Table III, each first place appearance preferably receives four times as much weight as a third place appearance. Each second place appearance receives 2 times as much weight as a third place appearance. Fourth through Sixth place appearances receive no weight. Thus an algorithm embodiment of the present invention can be constructed, as shown in Table IV.

TABLE IV

Site-A's "total weight" = (1*4)+(1*2)+(1*1) = 7;
Site-B's "total weight" = (1*4)+(0*2)+(2*1) = 6;
Site-C's "total weight" = (1*4)+(1*2)+(0*1) = 6;
Site-D's "total weight" = (2*4)+(0*2)+(2*1) = 10;
Site-E's "total weight" = (0*4)+(2*2)+(1*1) = 5; and
Site-F's "total weight" = (1*4)+(2*2)+(2*1) = 10.

There are several advantages in using such a method. The sites that do the best will generally receive more connections than other sites, but not too many of the connections. Any hand-offs that occur is preferably averaged across the top few sites, and such is made tunable by adjusting the static hand-off weighting. The sites that are

seen as poorly performing by all other sites will tend to receive fewer or no hand-offs. If every site is performing well, including WAN links, servers, etc., then its likely that each site will receive an equal distribution of traffic over time.

A calculated hand-off table, such as Table III, is principally used for DNS response ordering and "HTTP redirect" preference. It is not used when a TCP connection request comes to a VIP unless an "HTTP redirect" is called for.

When three or fewer sites are involved in a monitoring and hand-off exchange process, the poor granularity in the hand-off determination may be a problem. In such a case, there will not be enough throughput-data samples to accurately determine "best" versus "worst" sites, except in the most extreme of cases. Controls and tunable parameters within the switches should be included to mitigate this issue in such environments. A promising algorithm to use is a set of comparisons of the VIPCONNS to MAXCONNS ratios. A site that can accept the most connections will have a tendency to receive the most connections.

DSSP triggered updates preferably contain all of the information that a regular update has, but such are sent immediately from one switch to all other switches when the switch is (a) no longer able to communicate with a remote server, or (b) when the switch experiences a local resource constraint, such as all servers are at their respective MaxConns, no real servers are available for a VIP, etc.

To illustrate a DSSP-update example, a site-A has five peers sites B-F. Each site A-F runs two VIP's and are peered with every other site. For session hand-off distributed-server determinations, each site's switch computes an ordered hand-off table for each matching domain name for each remote VIP/Local VIP combination. Each switch communicates a VIP that represents "www.Alteon.com", and an entry will appear in a calculated hand-off table based on the test responsiveness of each VIP. For a given domain name, such as "www.alteon.com", an ordered hand-off table is preferably constructed by each switch. The hand-off table is thereafter consulted when the switch receives a domain name server request for the domain name the table is constructed for. Each switch will dynamically update the remote real server's weight based upon computed weight values, as illustrated in the Tables herein. When the domain name server request for "www.alteon.com" is received by any switch, it will respond with the IP-address that corresponds to the "next eligible"

remote server, based on the current weights. The VIP corresponding to distributed-server F will generally receive 25% of the requests. In other words, 25% of the time any switch receives a domain name server request, the switch will respond with distributed-server's VIP-address.

TABLE V: Ordered Hand-off Weight Table
DNS/HTTP Redir Hand-off Weights (with Traff Dist)

site	total weight	traffic disti	order
A	11	26%	weight-1
B	12	29%	weight-2
C	0	0%	weight-3
D	6	14%	weight-4
E	10	24%	weight-5
F	3	7%	weight-6

In the ordered hand-off table, Table V, site-C has a weight of zero. This VIP should never have any hand-off requests sent to it. In this example, sites A, B, and E will receive the majority of the hand-offs.

For session hand-off execution, when a switch receives a domain name server request for a domain name that it is hosting, it will respond with the appropriate IP-addresses of the switches that are load balancing those domains, based on hand-off weights, availability, etc. It is important to take into account the physical proximity when doing a hand-off. Generally, it is preferably best if users within a region are associated with servers in or near that region, unless the nearby server is down or overloaded. For example, let's say there are five sites that host content for "www.akeon.com" installed all over the world: San Jose (West-US); Atlanta (East-US), Ecuador (South America), Paris (France), and Tokyo (Japan). Users in Europe are preferably served by the Paris site, users in Chile are preferably served by the Ecuador site, etc. Having a user in Japan come all the way to the Atlanta site for content would waste bandwidth that many other users could have benefited from, and such service would directly result unnecessary response delays to the Japanese user.

It is therefore important for a switch to weigh-in to the final decision the geographic source of a user request prior to performing any session hand-off. When a switch receives a domain name server request for a domain that it is configured for, the switch should inspect the source IP-address of the request, and generally associate it with the IP-address blocks issued from IANA to the various regional registries. Table VI shows some of the address block allocations for the various regional registries, and their respective geographic domains.

TABLE VI

RIPE NCC - Europe Apr 97	063/8
ARIN Apr 97	064-095/8
RIPE NCC - Europe May 93	194/8
RIPE NCC - Europe May 93	195/8
RIPE NCC - Europe May 93	196/8
ARIN - North America May 93	199/8
ARIN - North America May 93	200/8
ARIN - Central and South America May 93	201/8
ARIN - Central and South America May 93	202/8
APNIC - Pacific Rim May 93	203/8
APNIC - Pacific Rim May 93	204/8
ARIN - North America Mar 94	205/8
ARIN - North America Mar 94	206/8
ARIN - North America Apr 95	207/8
ARIN - North America Nov 95	208/8
ARIN - North America Apr 96	209/8
ARIN - North America Jun 96	210/8
APNIC - Pacific Rim Jun 96	211/8
APNIC - Pacific Rim Jun 96	212/8
RIPE NCC - Europe Oct 97	213/8
ARIN - North America Apr 98	217/8

An extension of Table VI is preferably provided in a database form that can be accessed by each switch embodiment of the present invention. The source network

is preferably resolved to a 124-bit IP subnet depth. The database used is preferably derived from the IANA "WHOIS" database. Using such a table of information in the switch will allow the domain name server responder to make a rough geographic decision on the source of the domain name server request. If the domain name server request is 211.123.11.20, the requesting host is located somewhere in the Pacific Rim area, and should be pointed to a site that begins with either 203, 204,211, 212. The switch preferably uses this table of information during all domain name server responses if any of the distributed sites VIP's are on geographically diverse networks.

In a peer hand-off process, a switch receives domain name server lookup request for a particular VIP domain name. The switch examines the source IP address for the domain name server request, examines the user's IP-address, and determines if there is server site that is geographically close to that user. The switch examines an ordered hand-off table corresponding to the domain. The switch chooses a next remote server (or its own VIP) in line based on, (a) the remote server location compared to domain name server request source, (b) the remote servers weights, and (c) remote server that experienced the previous hand-off. The switch then sends the domain name server response back to client domain name server with the IP-addresses in an ordered list.

When the switch receives a "TCP SYN" to switch VIP, it either accepts packet or rejects the packet if the local VIP is overloaded. If rejected, the switch examines ordered hand-off table for this domain, and chooses a next remote server or its own VIP in line based on, (a) the remote servers location compared to domain name server request source, (b) the weights of each remote server, and (c) the remote server identified in a previous hand-off. The switch sends an "'HTTP redirect'" back to the client or drops the request, depending on load and availability of other sites.

When a switch issues a domain name server response, it will do so with a configurable domain name server TTL value, to ensure that downstream domain name server's do not cache the server switch's IP-address for too long a period of time.

For distributed load balancing parameters, each switch is preferably configured with switch-wide distributed SLB-parameters to recognize its distributed sites. For example, by a list of all the other switches' management IP-addresses.

Various tunable parameters are preferably included in embodiments of the present invention. Distributed sites with eight configurable distributed sites per switch, are configured with the remote switches' IP-addresses. Each of these sites can be recognized by a switch as a potential hand-off site where remote real servers (VIP's) exist. The distributed-server state protocol interval represents how often switches communicate regular DSSP updates. A range of 1-120 minutes is preferred with a default of one minute and may be turned off for individual sites. A domain name server TTL represents the TTL-value that is to be used when responding to domain name server requests. A range of 0-255 minutes is preferred with a default of one minute. For distributed SLB on/off controls, the ""HTTP redirect"" option can be used and set to "On/Off" with the default being "On," and also the "UseDNSRespond" option, which can be set to "On/Off," with the default being "On." Ordered Hand-off Weights (indexed 1-16), which can have a value of 1-64, to be taken into account while computing the ordered hand-off list.

Each hand-off weight index (1,2,3... 16) corresponds to a best-performing to a worst performing-site. Each index can have a statically configured weight that is preferably multiplied by the server switch's relative positions in the ordered hand-off list. If the ordered hand-off weight (OHW) index-1 is set to four, the best performing site will receive four-times the connections of a site with a weight of one. A typical configuration may be to set: OHW-1 to "6", OHW-2 to "4", OHW-3 to "2", and all others to "1". This will lead to the first, second and third best performing sites to receive six times, four times, and two times as many hand-offs compared to the rest of the server switches.

Fig. 3 represents a flowchart of a distributed-server web-balance method embodiment of the present invention, and is referred to herein by the general reference numeral 300. The method 300 begins with a step 302 in which a user request for a DNS-lookup has been received. Such request asks for a numeric IP-address that will respond with a particular web-based content and service. A step 304 determines what the geographic domain of the user is by inspecting the user IP-address included in the DNS-lookup query. A step 306 examines the available network sites and switches in or near the user's geographical area. A step 308 calculates the "best" virtual IP-server (VIP) that should be given the job of corresponding afterward with the user. What constitutes "best" depends on what

goals are being addressed. "Best" could be best overall system performance from the perspective of the user, the web-site, the backbone operator, the Internet Service Provider (ISP), cost, etc. A background process 310 continually monitors the health and performance of all the VIP's. A step 312 responds to the user's DNS-lookup request with the IP-address of the "best" VIP to service the user.

Although the present invention is described herein with reference to the preferred embodiment, one skilled in the art will readily appreciate that other applications may be substituted for those set forth herein without departing from the spirit and scope of the present invention. Accordingly, the present invention should only be limited by the Claims included below.

CLAIMS

1. A distributed load-balancing Internet server system for providing web-based content and services to be redundantly delivered to many clients, comprising:

a domain name system (DNS) server for receiving a DNS-lookup request from a network user for a conversion of a particular uniform resource locator (URL) for a domain host name to a numeric Internet Protocol (IP) address, wherein said network user exists in a particular geographical area that can be discerned from a user IP-address;

a plurality of web-server sites that are geographically diverse and accessible to said network user, wherein each duplicates another in its web-based content and services that relate to said particular URL;

a policy manager that monitors the health and response performance of each of the plurality of web-server sites, and that maintains a list of such ones of the plurality of web-server sites according to their individual accessibility and geographic location; and

a DNS-query to IP-address converter connected to receive said DNS-lookup request from said network user, and connected to consult the policy manager for a preferred one of the plurality of web-server sites to respond to such DNS-lookup request, and further connected to provide said network user with an IP-address of said preferred one of the plurality of web-server sites.

2. The system of claim 1, wherein:

each of the plurality of web-server sites corresponds to a virtual IP-address (VIP) and is physically located at a different place in the world.

3. The system of claim 1, wherein:

each of the plurality of web-server sites is able to off-load the others and operate in parallel to serve many simultaneous network users with diverse geographic loactions.

4. The system of claim 1, wherein:

the DNS-query to IP-address converter operates such that system-wide loads are balanced amongst each of the plurality of web-server sites.

5. The system of claim 1, wherein:

the policy manager further includes a response-time matrix and handoff table that maintains said list.

6. The system of claim 5, wherein:

the policy manager further includes a hand-off weight index that corresponds to a best-performing to a worst performing web-server site and a statistically configured weight that is multiplied by a relative positions in the ordered hand-off list of a server switch.

7. The system of claim 5, wherein:

the policy manager further includes an Internet topology awareness and a distributed SLB-state protocol that is capable of exchanging health, load and throughput information between web-server sites either periodically, or when triggered by a predefined event.

8. The system of claim 1, wherein:

the plurality of web-server sites includes a main-content site that provides all web-content and services for duplication by each other web-server site.

9. A method of providing web-based content and services from to many clients from load-balanced redundant sites in response to a single DNS-lookup request, the method comprising the steps of:

receiving at a domain name system (DNS) server a DNS-lookup request from a network user for a conversion of a particular uniform resource locator (URL) for a domain-host name to a numeric Internet Protocol (IP) address, wherein said network user exists in a particular geographical area that can be discerned from a user IP-address;

placing a plurality of web-server sites at geographically diverse locations that are accessible to said network user, wherein each web-server site duplicates another in its web-based content and services that relate to said particular URL;

monitoring with a policy manager the health and response performance of each of the plurality of web-server sites, and maintaining a list of such ones of the plurality of web-server sites according to their individual accessibility and geographic location; and

converting a DNS-query to IP-address in response to a receipt of said DNS-lookup request from said network user, and connecting to consult said policy manager for a preferred one of the plurality of web-server sites to respond to such DNS-

lookup request, and further connecting to provide said network user with an IP-address of said preferred one of the plurality of web-server sites.

10. The method of claim 9, wherein:

the step of placing a plurality of web-server sites is such that each of said plurality of web-server sites corresponds to a virtual IP-address (VIP) and is physically located at a different place in the world.

11. The method of claim 9, wherein:

the step of placing a plurality of web-server sites is such that each of the plurality of web-server sites is able to off-load the others and operate in parallel to serve many simultaneous network users with diverse geographic loactions.

12. The method of claim 9, wherein:

the step of converting is such that a DNS-query to IP-address converter operates such that system-wide loads are balanced amongst each of said plurality of web-server sites.

13. The method of claim 9, wherein:

the step of monitoring is such that said policy manager further includes a response-time matrix and handoff table that maintains said list.

14. The method of claim 13, wherein:

the step of monitoring is such that said policy manager further includes a hand-off weight index that corresponds to a best-performing to a worst performing web-server site and a statistically configured weight that is multiplied by a relative positions in the ordered hand-off list of a server switch.

15. The method of claim 13, wherein:

the step of monitoring is such that said policy manager further includes an Internet topology awareness and a distributed SLB-state protocol that is capable of exchanging health, load and throughput information between web-server sites either periodically, or when triggered by a predefined event.

16. The method of claim 9, wherein:

the step of placing is such that said plurality of web-server sites includes a main-content site that provides all web-content and services for duplication by each other web-server site.

Fig. 1

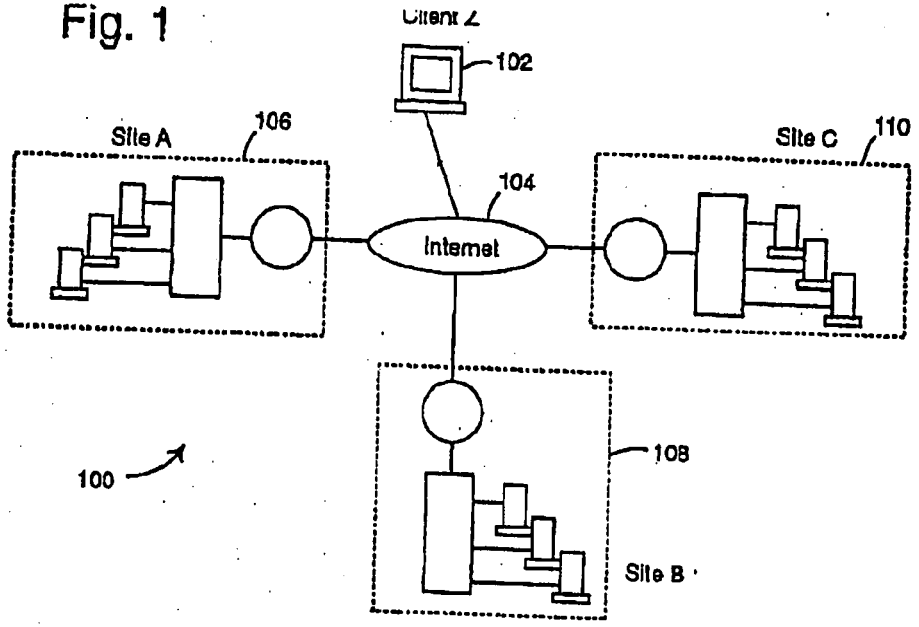


Fig. 2

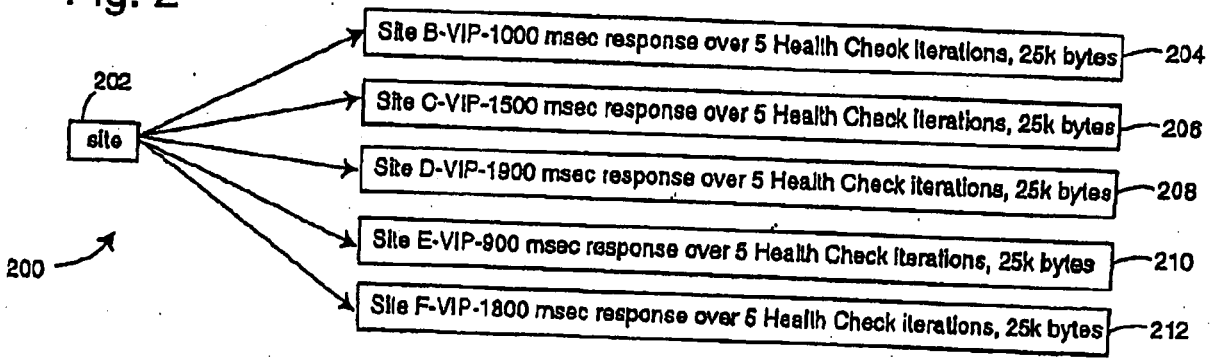
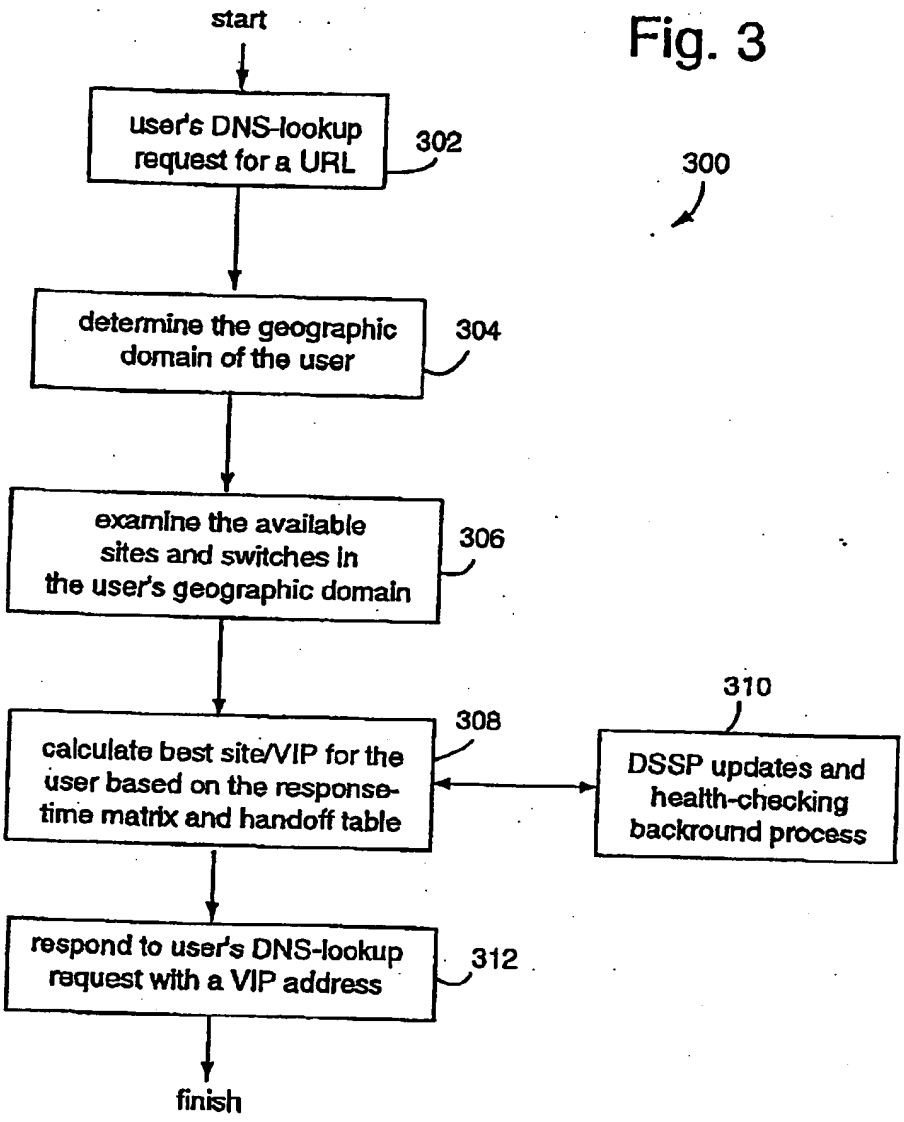


Fig. 3


start
user's DNS-lookup request for a URL
302
300
determine the geographic domain of the user
304
examine the available sites and switches in the user's geographic domain
306
calculate best site/VIP for the user based on the response-time matrix and handoff table
308
310
DSSP updates and health-checking backround process
respond to user's DNS-lookup request with a VIP address
312
finish

(19)00-315200 (P2000-31078

DISTRIBUTED LOAD-BALANCING INTERNET SERVERS

ABSTRACT

The actual site that serves the Web pages to a client in response to a URL domain name is automatically and transparently selected from a list of many switches each having identical data storage. In a peer hand-off process, a switch receives domain name server lookup request for a particular virtual Internet protocol (VIP) domain name. The switch examines the source IP-address for the domain name server request, examines the user's IP-address, and determines if there is server site that is geographically close to that user. The switch examines an ordered hand-off table corresponding to the domain. The switch chooses a next remote server (or its own VIP) in line based on, (a) the remote server location compared to domain name server request source, (b) the remote servers' weights, and (c) the remote server that experienced the previous hand-off. The switch then sends the domain name server response back to client domain name server with the IP-addresses in an ordered list.